Ataccama for Hadoop Improves Quality of Data and Productivity of Development

The Ataccama platform now supports Hadoop as a data source, and more importantly, developers can leverage the Ataccama visual data management tool to develop, deploy, and execute rich code on Hadoop nodes.

Toronto, Canada (April 2013) – Ataccama Corporation has announced its support for Hadoop in its market-proven Ataccama Data Quality Center (DQC). The company's aim is to continue providing its clients and the public with the current, already powerful DQC engine, while adding the ability to process the immense data loads of today's organizations. Hadoop-supported DQC allows companies to transition from their traditional data sources to new ones, leveraging Hadoop technology.

Big data is a highly overused term to describe a wide variety of current data problems. One of the most influential big data technologies is Hadoop—a general purpose, scalable computing platform. Among other capabilities, it allows for storage of any kind of data in a single system and supports a variety of subsequent needs for analytical computation. Hadoop offers cost-effective storage for all data, in addition to the data that is immediately needed, and offers an environment to run analytical tasks directly on stored data.

One of the key challenges with this new data storage/processing/analytical environment is that raw data stored in Hadoop might be of poor or unknown quality. Prior to any analytical task, inconsistencies must be dealt with (for example, invalid or missing values, duplicities, etc.). Several companies implementing a solution with Hadoop expressed their concern over the connected data quality, and were looking for a tool that would allow them to deal with quality issues directly in their Hadoop cluster—without a need to transport the data to and from the data quality engine. This is how the symbiosis of Ataccama DQC and Hadoop began.

In support of the push to combine the technologies, Ataccama director Michal Klaus explained, *"Some of the data management problems don't go away with Big Data, including quality of the data and need for master data. This is where Ataccama DQC for Hadoop can provide incredible value. Interestingly, in addition to being a great data quality engine, DQC has quite a few generic data integration features. To our surprise, we have learned that our early adopters are also leveraging DQC for Hadoop as a universal data manipulation tool for Hadoop data, because it is very easy to use and efficient. Now that is definitely exciting."*

Within the DQC application (see figure), the new *Hadoop Clusters* section presents the connected clusters and allows for simple drag and drop manipulation, as well as reading

common file types with the assistance of DQC Editor. The *Run on cluster* function performs the following key steps:

- Obtains the relevant DQC plan file and all other local source data (e.g. reference data not yet present in the Hadoop cluster) necessary for its proper running and copies the data to all available nodes in the connected cluster.
- Launches the plan in the massive parallel environment by using the Map Reduce paradigm.
- Stores the run result in the Hadoop cluster. These results can be used in subsequent processes, or displayed directly in DQC editor (in case of tasks producing a small aggregated result set, like profiling).

A major advantage of this new feature is that Ataccama DQC operates in exactly the same manner as with "standard data" projects. The only difference is that the application now allows for launching plans in a new type of environment, suitable for large datasets. Upon launching a run, an independent Ataccama engine operates on each connected cluster node to effectively perform all relevant tasks. And unlike other tools, Ataccama DQC for Hadoop is a native Hadoop application optimized for typical data integration and quality problems in a big data environment.

About Ataccama Software, s.r.o.

Ataccama Corporation is an international software company that prides itself in delivering cutting edge technology used for data quality, master data management, and data governance. Since 2007, leading financial, commercial, and government organizations choose Ataccama technologies for its proven ability to execute and deliver real business benefits.